histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data

Denis Schapiro[1,2,5], Hartland W Jackson[1,5], Swetha Raghuraman[1,4], Jana R Fischer[1], Vito R T Zanotelli[1,2], Daniel Schulz[1], Charlotte Giesen[1,4], Raúl Catena[1], Zsuzsanna Varga[3] & Bernd Bodenmiller[1]

Single-cell, spatially resolved omics analysis of tissues is poised to transform biomedical research and clinical practice. We have developed an open-source, computational histology topography cytometry analysis toolbox (histoCAT) to enable interactive, quantitative, and comprehensive exploration of individual cell phenotypes, cell–cell interactions, microenvironments, and morphological structures within intact tissues. We highlight the unique abilities of histoCAT through analysis of highly multiplexed mass cytometry images of human breast cancer tissues.

Technological advances in multiparametric analysis of single cells have allowed researchers to uncover the heterogeneity of cellular phenotypes and functional states concealed within population-based measurements[1–3]. Each cellular phenotype is defined by the interplay of its internal state and its environment, and tissue function is the output of these coordinated cell activities. Deregulation of intercellular communication is central to many diseases such as cancer[4]. Consequently, the ability to analyze single-cell functional states with spatial resolution is key to understanding normal tissue function and disease biology, and to the development of disease treatments.

Recent techniques such as FISSEQ[5], MERFISH[6], cycling immunofluorescence[7–9], multiplexed ion beam imaging (MIBI)[10], and imaging mass cytometry (IMC)[11] allow for single-cell, spatially resolved, highly multiplexed analysis of solid tissues and provide essential information on the distribution of transcripts, proteins, and protein modifications within single cells, microenvironments, and entire tissues[12]. Despite these experimental advances, no computational approach has been developed to enable comprehensive, quantitative, and interactive exploration of all levels of information within the data that result from spatially resolved, highly multiplexed tissue measurements. Current open-source tools that analyze image-linked data are typically focused on the analysis of cell lines imaged with fluorescence microscopy or basic tissue histology and are not geared for analysis of highly multiplexed measurements[13–15]. On the other hand, tools developed to perform analyses of nonimaging, multiplexed single-cell data such as that obtained using suspension-based mass cytometry do not exploit spatial information (**Supplementary Fig. 1**)[16,17].

In order to provide a complete picture of a tissue ecosystem, define molecular and spatial signatures necessary for analysis of tissue biology, and—in the case of disease—identify clinically relevant features, it is necessary to analyze and interrelate layers of information obtained from molecular measurements on cells, cell populations, cell–cell interactions, microenvironments, tissues, and experimental cohorts. Here we present an interactive computational platform called histoCAT that enables quantitative analysis of highly multiplexed and spatially resolved tissue measurements. histoCAT combines intuitive high-dimensional image visualization, state-of-the-art analysis methods for cell phenotype characterization, and novel algorithms for the comprehensive study of cell–cell interactions and the social networks of cells within complex tissues (**Fig. 1**). Thus, histoCAT provides a toolbox for investigating tissues in both healthy and diseased states.

In histoCAT, all single-cell information, including spatial features (**Fig. 1a**), is linked to the corresponding multiplex image enabling visualization of images and single-cell analysis in parallel (**Fig. 1a**,**b**). histoCAT uses a segmentation mask to extract single-cell data from images; such data include abundances of all measured markers for a cell and area of interest, spatial features like cell size and shape, and aspects of the cell's environment such as cell neighbors and cell crowding. This information is compiled into a flow cytometry standard format (.fcs) file for further analysis using histoCAT or other tools. 'Round-trip' analyses, those from a specific area of an image to data-set-wide analyses of single-cell phenotypes and their interactions and back to the visualization of unique cells on images, enable users to define and understand relevant cell populations and their spatial context in tissue. To enable quantitative and systematic analysis of all cell–cell interactions, we developed a novel algorithm to identify proximate cell–cell interactions that are present more frequently than expected by chance (**Fig. 1c**). Our algorithm enables determination of significant interactions and unique cell environments across entire data sets and within specific cohorts (**Fig. 1c**,**d**). These interactions can be represented in 'social networks' of cells and are complementary to interactions inferred from the presence of ligand receptor pairs[18].

To combine image-based spatial information and high-dimensional cytometry data, the histoCAT graphical user interface (GUI) is divided in two parallel sections for paired-image and

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [2]Life Science Zurich Graduate School, ETH Zurich and University of Zurich, Zurich, Switzerland. [3]Institute of Surgical Pathology, University Hospital Zurich, Zurich, Switzerland. [4]Present addresses: Institute of Cell Biology, ZMBE, University of Münster, Münster, Germany (S.R.) and F. Hoffmann-La Roche Ltd, Kaiseraugst, Switzerland (C.G.). [5]These authors contributed equally to this work. Correspondence should be addressed to B.B. (bernd.bodenmiller@imls.uzh.ch).

RECEIVED 7 FEBRUARY; ACCEPTED 7 JULY; PUBLISHED ONLINE 7 AUGUST 2017; DOI:10.1038/NMETH.4391

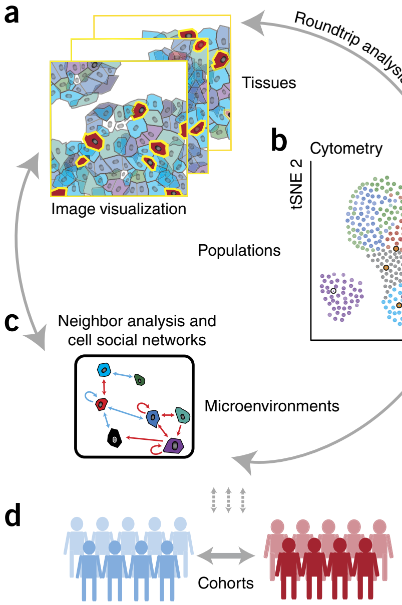

Figure 1 | Multiscale analysis of the tissue ecosystem. (**a**) Visualization of images, (**b**) cytometry analysis, and (**c**) analysis of neighbors and cellular interaction networks facilitate 'round-trip' analysis through layers of information. (**d**) Experimental cohorts can be compared and contrasted using molecular, cellular, and spatial signatures.

cytometry analysis (**Supplementary Note 1**). In the image visualization section of histoCAT, high-dimensional images as well as cell masks, single-cell marker quantification, and cell identification labels can be visualized. In the analysis section of histoCAT, image-derived marker quantification and spatial features of single-cell data are extracted for each image, combined (**Fig. 2a**), and visualized using multidimensional reduction tools such as tSNE[16] (**Fig. 2a**,**b**), scatter plots (**Supplementary Fig. 2**), histograms, box plots, and other visualizations (**Supplementary Note 1**).

To demonstrate the potential of histoCAT-powered analyses, we investigated cellular phenotypes and microenvironments of human breast cancer as visualized by IMC. By pairing classic immunohistochemistry staining, high-resolution tissue laser ablation, and mass cytometry, IMC can measure abundances of more than 40 unique metal-isotope-labeled tissue-bound antibodies simultaneously at a resolution comparable to that of fluorescence microscopy[11]. Here we analyzed images collected from 49 diverse breast cancer samples and three matched normal tissues, as well as from an additional six normal breast tissue samples. Tissues were stained with an antibody panel tailored to identify cell lineages and to detect signaling pathway activation, proliferation, apoptosis, and clinical markers (**Supplementary Tables 1** and **2**).

To gain a tissue-wide overview of cell phenotypes present in a given image set, we have incorporated two approaches into histoCAT. The first approach is supervised and based on tSNE[16], a data dimensionality reduction method that projects high-dimension cell data into two dimensions, grouping similar cells (as shown for image analysis of all samples in **Fig. 2a**,**b**). On the tSNE map in histoCAT, expression of individual markers can be highlighted using color scales, manually gated, and annotated for cell phenotype (**Supplementary Note 1**). The second approach is based on the unsupervised clustering algorithm PhenoGraph[19].

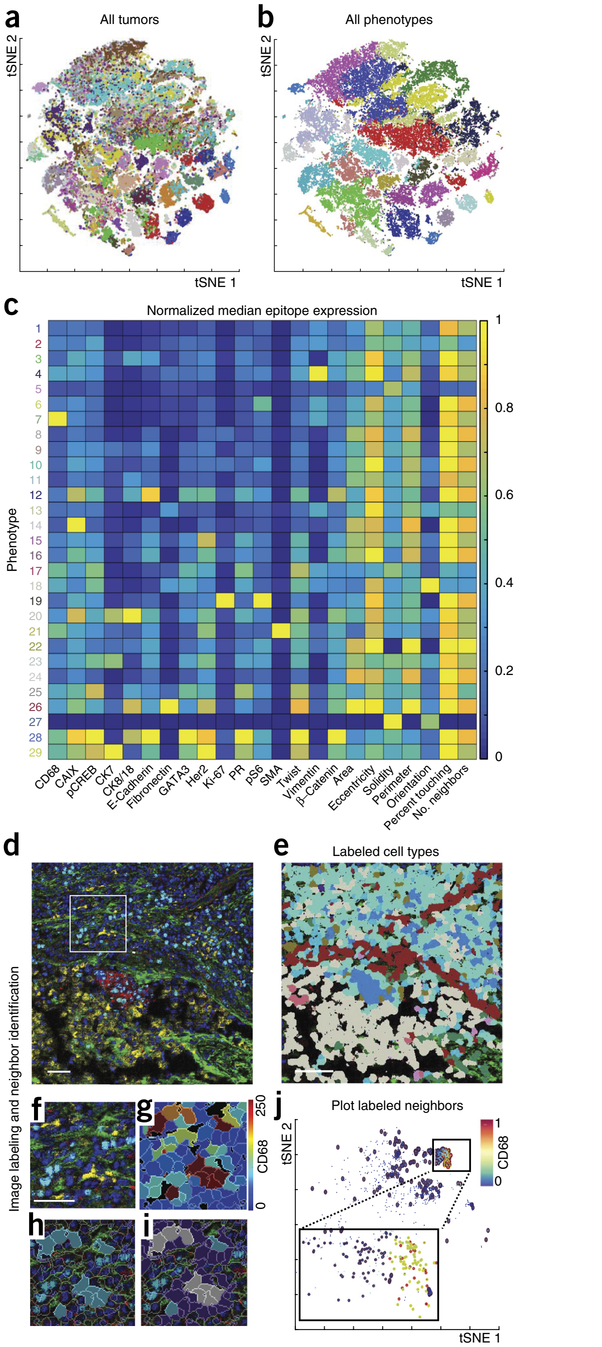

Figure 2 | Round-trip analysis of unique cell types in high-dimension images of breast cancer. (**a**) Entire breast cancer data set visualized in one t-SNE plot. Distinct colors distinguish cells from each source image. (**b**) PhenoGraph defines complex cell phenotypes based on marker expression and enables labeling of cell phenotype clusters on a t-SNE plot. (**c**) Cell phenotypes shown as heatmaps. (**d**) Cells of a subpopulation of interest can be identified. In this example red, E-cadherin; green, fibronectin; blue, histone H3; cyan, Ki-67; magenta, cytokeratin 7; yellow, CD68. Scale bar, 100 µm. (**e**) Single cells colored according to the identified phenotypes within the context of the tissue microenvironment on their original image. Non-cell tissue is not labeled. Images can be visualized using (**f**) pseudocolor or by (**g**) a heatmap representing the intensity of a marker in each cell. (**h**) Cells of interest can be highlighted on the image (turquoise), and neighboring cells (purple or gray if representing both subpopulations) within a defined pixel range can also be identified and highlighted on (**i**) the image or (**j**) the t-SNE analysis plot of the individual image (red, cell of interest; blue, neighbor; yellow, both subpopulations).

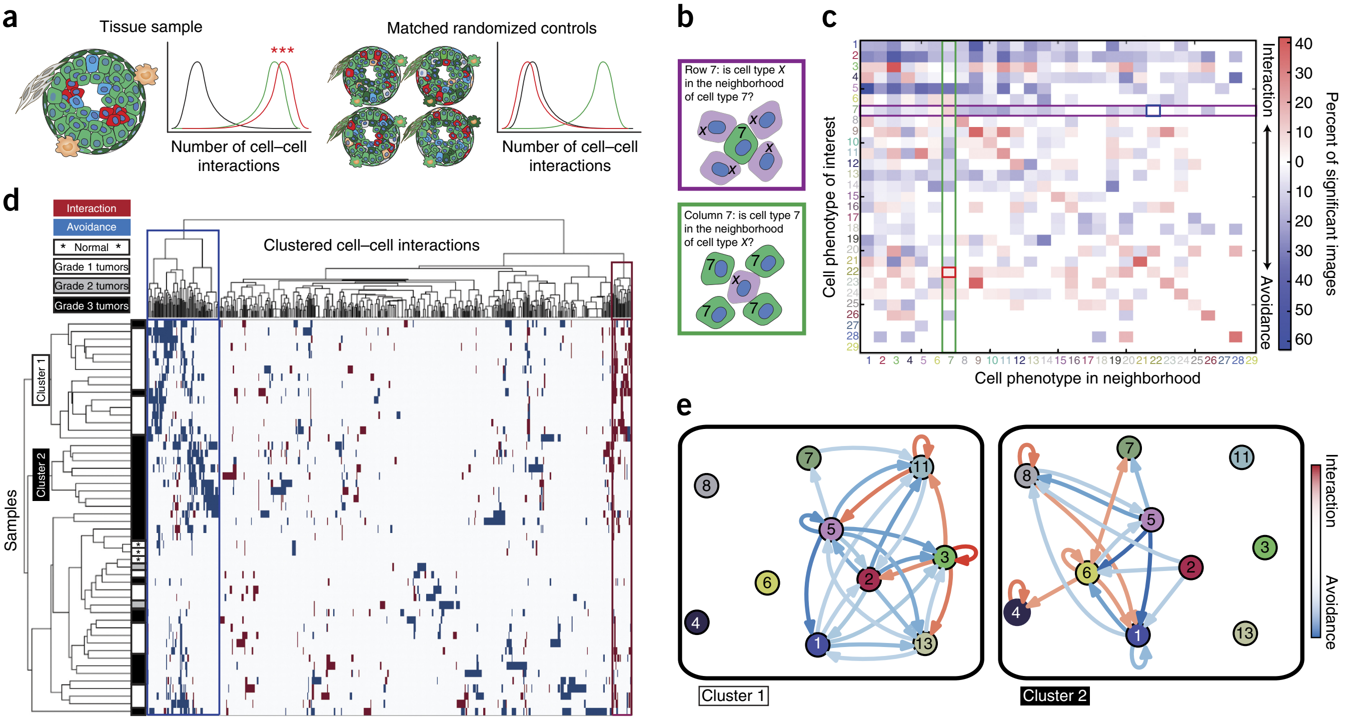

Figure 3 | Neighborhood analysis of breast cancer cell phenotypes. (**a**) Schematic of neighbor analysis. Number of interactions between abundant green cells (green line), between rare clustered red cells (red line), and between abundant green cells and rare red cells (black line). (**b**) Schematic depicting directional aspects of neighbor interactions visualized in the heatmap. Rows visualize the significance of all cell types surrounding a cell type of interest. Columns visualize the significance of the cell type of interest surrounding other cell types. White represents an interaction prevalence of less than 10%. (**c**) All interactions present in 49 breast tumor images and three matched normal tissue images are represented as a heatmap in which the cell type in the row is significantly neighbored (red) or avoided (blue) by the cell type in the column. Significance was determined by permutation test ($P < 0.01$). Highlighted squares indicate an example of a directional interaction. (**d**) Agglomerative clustering of all samples and cell–cell interactions according to the presence of significant ($P < 0.01$) phenotype interaction (red) or avoidance (blue). White represents interactions that are not present or not significant. (**e**) Cell social interaction network graphs representing the interactions of PhenoGraph-defined cell phenotypes in cluster 1 and cluster 2 tumors. Circle color corresponds to PhenoGraph cluster. Red arrows indicate interaction and blue arrows avoidance, and intensities of the line color indicate significance.

In the breast cancer and normal tissue samples, PhenoGraph identified 29 phenotype clusters (PC) shared across images and clinical subgroups, and these clusters were then visualized on a tSNE map (**Fig. 2c**, **Supplementary Fig. 3**, and **Supplementary Data Set 1**). These cell phenotypes are present at different frequencies and were characterized by specific epitopes (e.g., vimentin, PC 4; and CD68, PC 7) and combinations of markers (e.g., proliferative Ki-67$^+$ and phospho-S6$^+$ PCs 8, 10, and 19) (**Fig. 2c** and **Supplementary Fig. 3**). Marker intensities and cell populations can be linked back to their source images and visualized within the context of their multicellular environments (**Fig. 2d**,**e**).

Tumor-associated macrophages (TAMs) can drive or hinder tumor progression and are therefore highly attractive biomarkers and drug targets[20]. To gain a deeper understanding of TAMs and their neighborhoods, we inspected PC 7 with high CD68 signal suggestive of macrophage identity (**Fig. 2c** and **Supplementary Fig. 2b**). It is also possible to select potential macrophages by gating for CD68 expression on the tSNE plot (**Supplementary Note 1**). CD68 epitope expression was visualized in images by color (**Fig. 2d**,**f**, yellow) and heatmapped for each segmented cell (**Fig. 2g**). histoCAT also allows CD68$^+$ cells selected from a plot to be highlighted on source images in their original tissue context (**Fig. 2h**).

histoCAT has two neighborhood functions that enable investigation of the microenvironment. The first function is user guided and returns a subpopulation of cells touching or proximal to a cell population of interest for visualization on images (**Fig. 2i**) or for downstream analysis (**Fig. 2j**). This analysis performed on the 49 tumor-derived images and three matched normal tissues showed that distinct proliferative (Ki-67$^+$, phospho-S6$^+$) and hypoxic (carbonic anhydrase IX$^+$) epithelial tumor cells neighbor CD68$^+$ cells (**Supplementary Fig. 4**). The second neighborhood function enables the unbiased and systematic study of all cell–cell interactions present in a tissue or all tissues of a sample cohort by using a permutation test to compare the number of interactions between all cell types in a given image to that of a matched control containing randomized cell phenotypes (**Fig. 3a**). This approach determines the significance of cell–cell interactions and reveals enrichments or depletions in cell–cell interactions that are indicative of cellular organization. The significance of a neighboring interaction between each pair of phenotypes is visualized as a heatmap in which rows represent the neighborhood of a cell phenotype of interest and columns the enrichment or depletion of a cell in other neighborhoods (**Fig. 3b**,**c** and **Supplementary Data Set 2**).

We validated this strategy (see Online Methods; **Supplementary Notes 2** and **3**) on synthetic data (**Supplementary Fig. 5**), investigated its robustness to variations in cell segmentation (**Supplementary Fig. 6**), and showed that the algorithm identified known luminal–basal cell interactions in healthy mammary ducts and alveoli (**Supplementary Fig. 7**). Neighborhood analyses of images obtained from 49 breast cancer samples and three matched normal tissues identified cell phenotypes that surround or are surrounded by another cell phenotype (**Fig. 3b**,**c** and **Supplementary Fig. 8**). For TAMs (PC 7), unsupervised neighborhood analysis revealed that they surround multiple cell phenotypes (**Fig. 2e**; **Fig. 3c**, column 7 rows 6, 8, 22, 23, 24) identifying key TAM-interacting cells within all neighbor interactions (**Supplementary Fig. 4j**). Relationships and cellular crosstalk between CD68$^+$ cells and other cells, including phospho-S6$^+$/vimentin$^+$ stromal cells (PC 6) and E-cadherin$^+$/phospho-S6$^+$/Twist$^+$ tumor cells (PC 22), were identified, suggestive of distinct tumor microenvironments for future study.

To identify cellular landscapes, we clustered images based on all significant cell–cell interactions ($P \leq 0.01$, **Fig. 3d**). Distinct subgroups enriched in grade 1 and grade 3 tumors became apparent as well as a mixed-grade subgroup. To facilitate the visualization and comparison of cell–cell interactions over large data sets, we used our permutation-based neighborhood algorithm to identify social networks of cells that are specific to tumor grade (**Fig. 3e**). Separation of the grade 1 tumor subgroup was driven by tumor cell phenotypes (PCs 3, 9, 11) that interact with surrounding stromal cells (PCs 1, 5, 13), a cellular organization that reflects tissue tubularity, a major pathology grading criteria (**Fig. 3d**,**e**; **Supplementary Figs. 8a**,**b** and **9**). The more advanced lesions of grade 3 samples contained hypoxic cells (PC 14), interacting proliferative cells (PCs 8, 10, 19), and interactions with 'active' stroma (phospho-S6$^+$ and vimentin$^+$, PC 6) including macrophages (PC 7) (**Fig. 3d**,**e**; **Supplementary Figs. 8c**,**d** and **9**). Clustering of images based on significant cell interactions ($P \leq 0.01$), defined groups of tissues (in this case, cancer samples) that have similar organization and revealed pathology-grade-associated cellular ecosystems that may distinguish unique disease states.

By combining cytometry, image analysis, and novel algorithms for cell–cell interaction network analysis, histoCAT is able to define complex cell types using multiplexed measurements and spatial features as parameters and to elucidate patterns of cellular interactions within heterogeneous tissues. The use of 'round-trip' analyses between single-cell data and source images using machine learning and community-finding algorithms within an intuitive user interface will enhance our understanding of tissue structure at the cellular level. Combined with focused, hypothesis-driven data sets, future investigations of multiplexed imaging cytometry data using histoCAT could reveal cell types and cell interactions that drive disease. histoCAT is open source; we invite the community to further develop this toolbox for the analysis of next-generation imaging and pathology data.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank the Bodenmiller lab for support and fruitful discussions. Thank you to open-source softwares like "cyt", "CellProfiler," and many others. This work was supported by the Swiss National Science Foundation (SNSF) R'Equip grant 316030-139220, an SNSF Assistant Professorship grant PP00P3-144874, a Swiss Cancer League grant, the PhosphonetPPM and MetastasiX SystemsX grants, and funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. 336921. D. Schapiro was supported by the Forschungskredit of the University of Zurich, grant FK-74419-01-01, and the BioEntrepreneur-Fellowship of the University of Zurich, reference no. BIOEF-17-001. H.W.J. and D. Schulz are supported by European Molecular Biology Organization (EMBO) Long Term Fellowships cofunded by the European Commission (LTFCOFUND2013 and 2014), grants ALTF-711 2015 and ALTF-970 2014, respectively. H.W.J. was also supported by a Transition Postdoc Fellowship from the Swiss SystemsX.ch initiative ref. 2015/344, evaluated by the Swiss National Science Foundation.

AUTHOR CONTRIBUTIONS

D. Schapiro, H.W.J., and B.B. conceived of the project and software. H.W.J., C.G., and R.C. collected samples and validated antibodies. Z.V. assembled, classified, and provided tumor samples. H.W.J. completed the staining and image acquisition. D. Schapiro, S.R., and J.R.F. wrote the code. D. Schapiro, H.W.J., and D. Schulz tested software on multiple data sources. D. Schapiro, H.W.J., and V.R.T.Z. analyzed the images and single-cell data. D. Schapiro, H.W.J., and B.B. prepared the figures and wrote the manuscript. B.B. directed the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Tirosh, I. *et al. Science* **352**, 189–196 (2016).
2. Bendall, S.C. *et al. Science* **332**, 687–696 (2011).
3. Bodenmiller, B. *et al. Nat. Biotechnol.* **30**, 858–867 (2012).
4. Tabassum, D.P. & Polyak, K. *Nat. Rev. Cancer* **15**, 473–483 (2015).
5. Lee, J.H. *et al. Science* **343**, 1360–1363 (2014).
6. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. *Science* **348**, aaa6090 (2015).
7. Lin, J.-R., Fallahi-Sichani, M. & Sorger, P.K. *Nat. Commun.* **6**, 8390 (2015).
8. Schubert, W. *et al. Nat. Biotechnol.* **24**, 1270–1278 (2006).
9. Gerdes, M.J. *et al. Proc. Natl. Acad. Sci. USA* **110**, 11982–11987 (2013).
10. Angelo, M. *et al. Nat. Med.* **20**, 436–442 (2014).
11. Giesen, C. *et al. Nat. Methods* **11**, 417–422 (2014).
12. Bodenmiller, B. *Cell Syst.* **2**, 225–238 (2016).
13. Ding, H., Wang, C., Huang, K. & Machiraju, R. *BMC Bioinformatics* **16**, S10 (2015).
14. Beck, A.H. *et al. Sci. Transl. Med.* **3**, 108ra113 (2011).
15. Jones, T.R. *et al. BMC Bioinformatics* **9**, 482 (2008).
16. Amir, A.D. *et al. Nat. Biotechnol.* **31**, 545–552 (2013).
17. Shekhar, K., Brodin, P., Davis, M.M. & Chakraborty, A.K. *Proc. Natl. Acad. Sci. USA* **111**, 202–207 (2014).
18. Rieckmann, J.C. *et al. Nat. Immunol.* **18**, 583–593 (2017).
19. Levine, J.H. *et al. Cell* **162**, 184–197 (2015).
20. Ostuni, R., Kratochvill, F., Murray, P.J. & Natoli, G. *Trends Immunol.* **36**, 229–239 (2015).

ONLINE METHODS

Preparation and staining of breast cancer tissue specimens. Formalin-fixed paraffin-embedded tissue samples from patients treated at the University Hospital Zurich between 1991 and 2005 were retrieved from the archives of the Institute of Surgical Pathology. This project was approved by the local Commission of Ethics (ref. no. StV 12-2005). H&E-stained sections of all tumors were re-evaluated by a pathologist for their suitability for tissue microarray construction before array construction as previously described[21].

Tissues were stained as previously described[11]. The antibody panel with clone information is shown in **Supplementary Table 1**. Briefly, tissue sections were dewaxed overnight in xylene and rehydrated in a graded series of alcohol (ethanol:deionized water 100:0, 90:10, 80:20, 70:30, 50:50, 0:100; 5 min each). Heat-induced epitope retrieval was conducted in a water bath at 95 °C in Tris-EDTA buffer at pH 9 for 20 min. After immediate cooling, the microarrays were blocked with 3% BSA in TBS for 1 h. Samples were incubated overnight at 4 °C in primary antibody at 7.5 g/L diluted in TBS/0.1% Triton X-100/1% BSA (clones in **Supplementary Table 1** and **Supplementary Note 4**). Panel design and antibody database management was done in AirLab[22]. Samples were then washed twice with TBS/0.1% Triton X-100 and twice with TBS, and they were dried before imaging mass cytometry measurement.

Imaging mass cytometry. Antibody staining of tissue sections was quantified through the combination of laser ablation using a modified ArF excimer GeoLas C laser system (Coherent) to ablate tissues in a rastered pattern at 20 Hz and to direct aerosol transportation of the sample to a CyTOF mass cytometer (Fluidigm) as previously described[23]. All raw data processing was performed using in-house Matlab routines as previously described and provided[11].

Segmentation. Segmentation was performed using Ilastik 1.1.9 (ref. 24) and CellProfiler 2.1.1 (ref. 15). Ilastik was used to classify pixels into three classes (nuclei, membrane, and background) and to generate probability maps. CellProfiler was used to segment probability maps to generate segmentation masks. A combination of channels was used to classify the background and membrane[25]. These masks were combined with the individual tiff files to extract single-cell information from each individual image.

To quantitatively assess the segmentation quality, we used four intuitive segmentation constraints for a segmentation score[25]: (i) mask should overlap with membrane signal; (ii) mask should not overlap with nuclei signal; (iii) segmented cells should contain maximally one nucleus; (iv) mask should approximate the expected number of cells based on cell radius. We also analyzed the effect of segmentation by using segmentation masks generated by five independent users on three test images (**Supplementary Fig. 6**). The complexities and difficulties of image segmentation are discussed in **Supplementary Note 3**.

Single-cell feature extraction. histoCAT uses Matlab's regionprops function to extract shape and pixel value measurements. Additionally, by step-wise pixel expansion, histoCAT creates a network of neighbors surrounding each cell at a range of defined distances. In most cases, expansion in the range of 1 to 6 pixels was chosen. Cells were expanded using a rectangular membrane shape. All cells within the defined range were considered neighbors. The distances between centroids were used to define cell–cell distances. The number of neighbors and the percent of membrane in contact with a neighboring cell were both determined with modified CellProfiler 1.0 modules[15].

Data transformation. Raw measurements were used for the presented data. histoCAT also offers arcsinh transformation with a variable cofactor input.

Normalization. All images were segmented, and single-cell measurements were extracted from all available channels using the mean pixel values for each segmented cell. The presented data were not normalized, but histoCAT features Z-score normalization across all samples or across a subgroup as a module. We used 99th-percentile normalized data for t-SNE and PhenoGraph as suggested[16,19]. Heatmaps were visualized using scaled values from 0–1 for individual columns in analysis plots as well as on images for individual masks.

Histology topography cytometry analysis toolbox. histoCAT can be downloaded either as a Matlab 2014b app or as a stand-alone application for Mac OS12 (**Supplementary Software 1**), Windows 7 (**Supplementary Software 2**), and Windows 10 (**Supplementary Software 3**) from https://github.com/BodenmillerGroup/histoCAT. Documentation, user manual, and development versions of histoCAT can also be found on the project page. All modules, if not differently stated, were written in Matlab 2014b, and the GUI was designed in Matlab 2014b using Matlab's GUI development environment (GUIDE). histoCAT is built modularly to enable addition of new features without the need for changes to the existing structure. In general, features in histoCAT must include only two basic scripts, callback from the GUI and the script executing the function. The main functions are not linked to the GUI and can be run independently.

All data necessary to perform any function for the current session can be retrieved from the GUI handles or included manually without the GUI. Throughout a session, the data are kept in the fcs-format structure. There is one main matrix containing a column for each channel and a row for each individual cell of each image. This matrix is continuously updated during the session and will therefore also contain the custom gates and channels. The corresponding channel names for each image are saved in a cell array. All individual tiff files and corresponding masks are stored in a multidimensional matrix structure.

Barnes–Hut t-SNE. We used the Barnes–Hut t-SNE implementation in histoCAT[16]. Data were 99th-percentile normalized before the analysis, and we used the default t-SNE parameters (initial dimensions, 110; perplexity, 30; theta, 0.5). The random seeds for the individual runs can be recorded.

PhenoGraph. PhenoGraph version 0.2 was used[19]. Data were 99th-percentile normalized before the analysis, and default parameters with nearest neighbors of 75 were used. This parameter was chosen based on prior knowledge of the underlying cell types. Lower values for nearest neighbors result in an overclustering and higher values an underclustering. The random seeds for the individual runs can be recorded.

doi:10.1038/nmeth.4391

Neighborhood analysis, permutation test. The neighborhood analysis uses basic statistical methods to find significantly enriched interactions between or within cell phenotypes. First, cells are manually or automatically classified. Manual classification can be done by manual gating on biaxial/t-SNE plots. Automatic classification uses the PhenoGraph[19] algorithm, which consistently performs well for data sets with multiple cell populations[26].

Once classified, pairwise interactions at a user-defined distance (6 pixels) between and within cell phenotypes are calculated for each single cell with its neighbors. A neighbor is defined as a cell within the pixel distance selected during the loading process. Pairwise interactions between and within cell phenotypes are compared to a random distribution using two individual one-tailed permutation tests (equation (1)). These provide two *P* values which correspond to either interaction or avoidance. These *P* values represent the likelihood of a neighborhood interaction being enriched or depleted in comparison to a randomized version of the same tissue. The comparison to a matched randomized tissue for every individual image controls for both the distinct connectivity and the specific cell types in that tissue. Equation (1) describes our approach using a permutation test with Monte Carlo sampling. We run this test twice to calculate the *P* value for each tail.

$$P = \Sigma(\text{mean(permutations)} \geq (\leq) \text{mean(real data))} / \text{ no. permutations} + 1 \qquad (1)$$

Agglomerative hierarchical clustering using inner squared distance (minimum variance algorithm employing Ward's method) as linkage criteria and Euclidean distance as the distance metric is applied to display similar interaction signatures in a dendrogram of all samples.

To facilitate the visualization and comparison of cell–cell interactions and cellular social networks we have used cell interaction network graphs representing the interactions of phenotypes. A connection is only visualized if the interaction or avoidance is significant ($P < 0.01$) in a least 30% of the grouped samples and the cell phenotypes are simultaneously present in at least 90% of the grouped samples.

Analyses of synthetic data (**Supplementary Fig. 5**), variations in single-cell masks (**Supplementary Fig. 6**), and organized normal breast tissue (**Supplementary Fig. 7**) validated and highlighted the robustness of our neighborhood analysis. In the case of rare cell populations (i.e., low cell number), our algorithm is able to detect significantly enriched interactions (**Supplementary Notes 2** and **3**). In contrast, avoidance in rare cell populations will rarely be detected on account of the minimal interactions among the rare cell population in the random shuffled control. The absolute cell number or the size of the image does not have a direct effect on the neighborhood analysis, as described above; only the frequencies of cell types present and the relative quantities are important. The expected average interactions can be approximated by equation (2), which includes the amount of cell type A multiplied by the fraction of cell type B and by the average connectivity of the image. If the average connectivity in an image is high, the expected interactions are high and vice versa.

$$\text{expected}_{\text{interactions}} = \text{Cell A} \times \frac{\text{no.Cell B}}{\text{no.all Cells}} \times \text{average}_{\text{connectivity}} \qquad (2)$$

A **Life Sciences Reporting Summary** for this paper is available.

Data availability statement. All software and code that produced the findings of the study, including Mac OS12 (**Supplementary Software 1**), Windows 7 (**Supplementary Software 2**), and Windows 10 (**Supplementary Software 3**) versions of histoCAT are available at https://github.com/BodenmillerGroup/histoCAT and http://www.bodenmillerlab.org/research-2/histoCAT/. All raw data, histoCAT sessions, and interactive graphs that support the findings of the study are available at http://www.bodenmillerlab.org/research-2/histoCAT/. Source data for **Figure 3** is available online.

21. Kononen, J. *et al. Nat. Med.* **4**, 844–847 (1998).
22. Catena, R., Özcan, A., Jacobs, A., Chevrier, S. & Bodenmiller, B. *Genome Biol.* **17**, 142 (2016).
23. Wang, H.A.O. *et al. Anal. Chem.* **85**, 10107–10116 (2013).
24. Sommer, C., Straehle, C., Köthe, U. & Hamprecht, F.A. in Proc. 2011 8th IEEE International Symposium on. *Biomedical Imaging: From Nano to Macro* 230–233 (IEEE, 2011).
25. Schüffler, P.J. *et al. Cytometry A* **87**, 936–942 (2015).
26. Weber, L.M. & Robinson, M.D. *Cytometry A* **89**, 1084–1096 (2016).